SOCR: A Handwritten Data Form Producing and Reading System

Hanchuan Peng and Qiang Gan

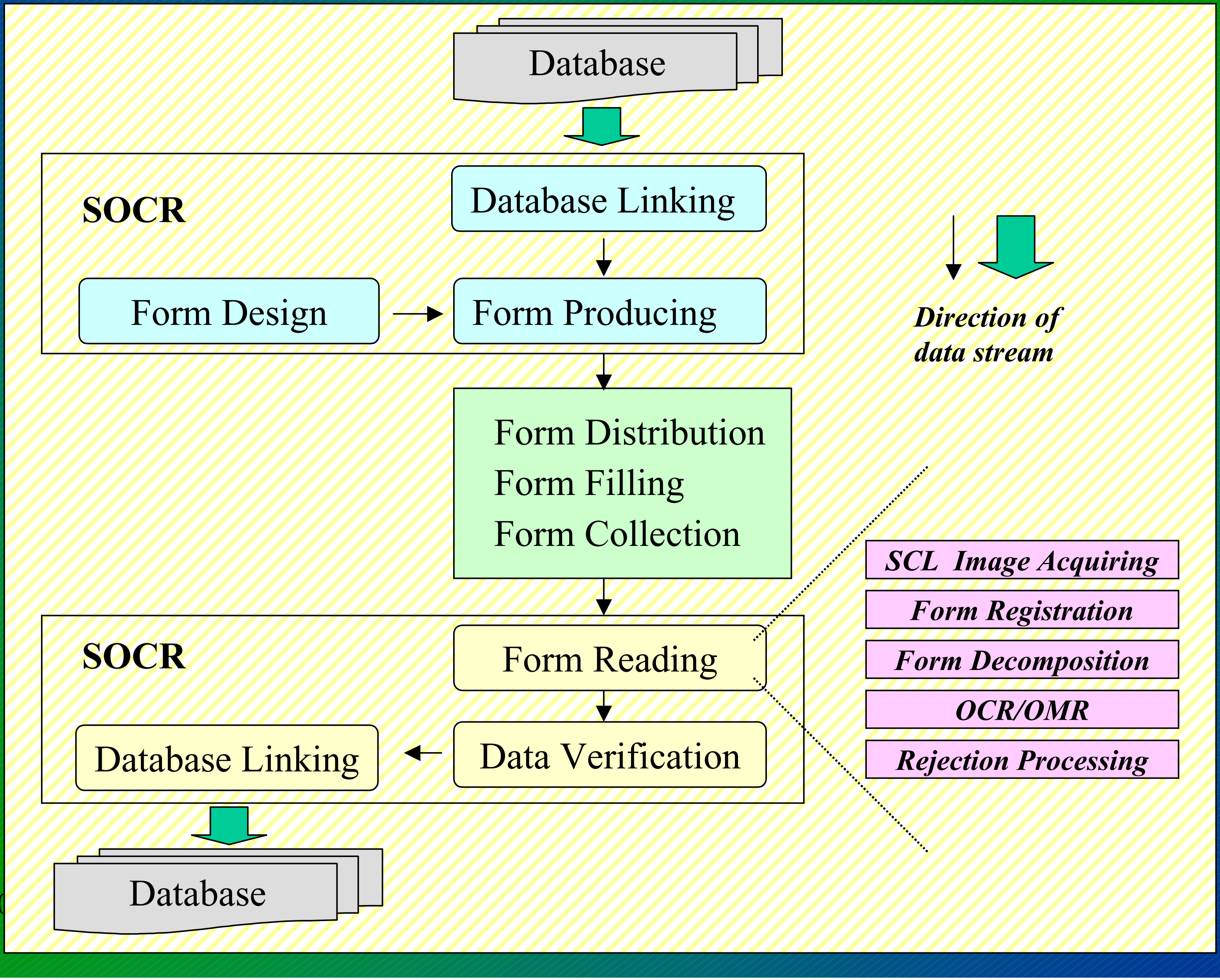
Department of Biomedical Engineering,
Southeast University,
Nanjing, 210096, China.
Email: phc@seu.edu.cn

2000-Jan-10

Abstract

Document analysis and recognition software, especially form reader, is greatly required in office automation. This paper serves as a primer to the SOCR (version 1.03), which is a software package for handwritten data form producing and reading. This package integrates techniques of database linking, form producing, form reading, and data verification. With SOCR, data forms can be easily created and the form data, including loosely constrained handwritten numerals and symbols, can be read into databases at high accuracy and high speed, and in a human-compatible manner. This software package has been applied to producing and reading handwritten student score forms in several universities and tax forms in several cities.

Main Structure of the System



Database

SOCR

Database Linking

Form Design

Form Producing

Form Distribution
Form Filling
Form Collection

SOCR

Form Reading

Data Verification

Database Linking

Database

Direction of data stream

SCL Image Acquiring

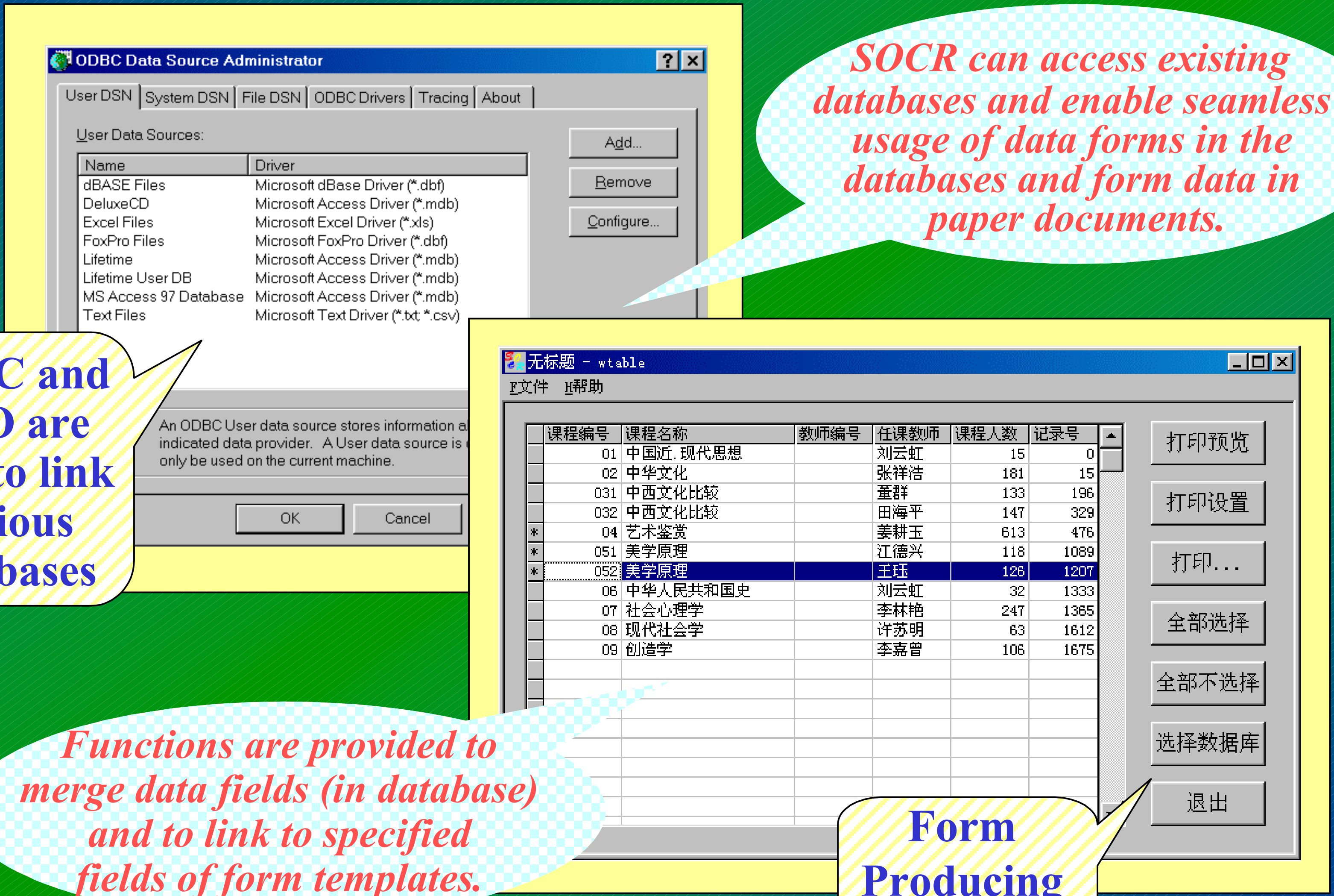
Form Registration

Form Decomposition

OCR/OMR

Rejection Processing

Database Linking and Form Producing



ODBC Data Source Administrator

User DSN | System DSN | File DSN | ODBC Drivers | Tracing | About

User Data Sources:

Name	Driver
dBASE Files	Microsoft dBase Driver (*.dbf)
DeluxeCD	Microsoft Access Driver (*.mdb)
Excel Files	Microsoft Excel Driver (*.xls)
FoxPro Files	Microsoft FoxPro Driver (*.dbf)
Lifetime	Microsoft Access Driver (*.mdb)
Lifetime User DB	Microsoft Access Driver (*.mdb)
MS Access 97 Database	Microsoft Access Driver (*.mdb)
Text Files	Microsoft Text Driver (*.txt; *.csv)

An ODBC User data source stores information a
indicated data provider. A User data source is
only be used on the current machine.

无标题 - wtable

	课程编号	课程名称	教师编号	任课教师	课程人数	记录号
	01	中国近.现代思想		刘云虹	15	0
	02	中华文化		张祥浩	181	15
	031	中西文化比较		董群	133	196
	032	中西文化比较		田海平	147	329
*	04	艺术鉴赏		姜耕玉	613	476
*	051	美学原理		江德兴	118	1089
*	052	美学原理		王珏	126	1207
	06	中华人民共和国史		刘云虹	32	1333
	07	社会心理学		李林艳	247	1365
	08	现代社会学		许苏明	63	1612
	09	创造学		李嘉曾	106	1675

SOCR can access existing databases and enable seamless usage of data forms in the databases and form data in paper documents.

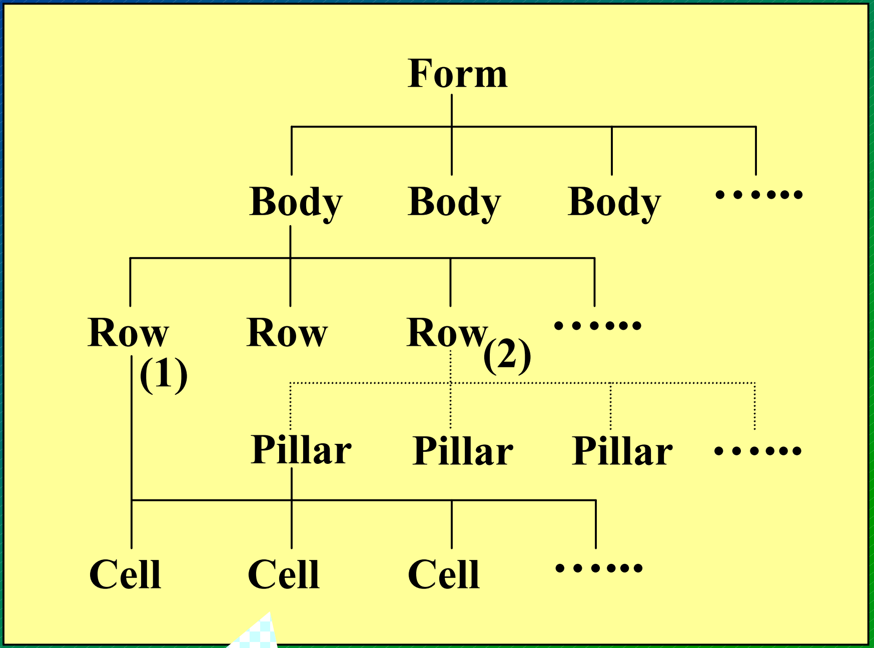
ODBC and DAO are used to link various databases

Functions are provided to merge data fields (in database) and to link to specified fields of form templates.

Form Producing Wizard

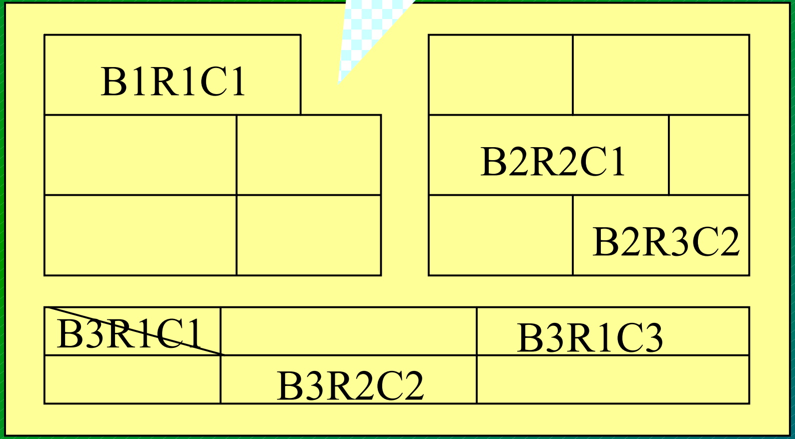
2000-Jan-10

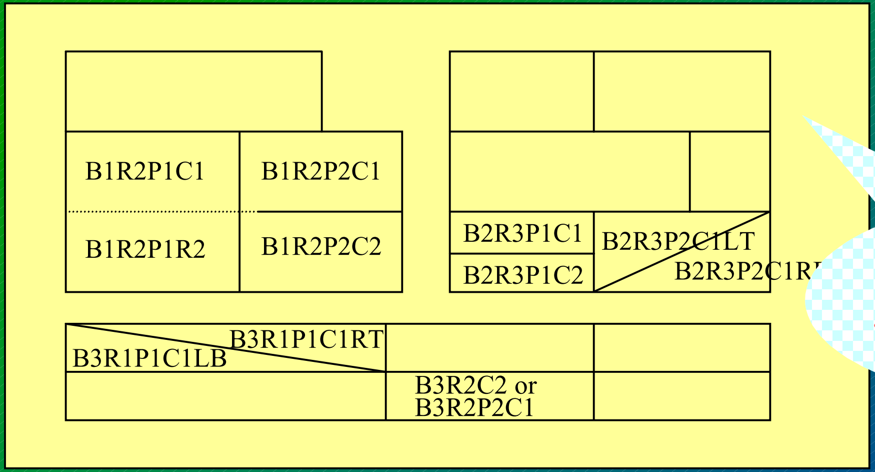
Form Data Structure



The hierarchy tree, where the level of "Pillar" is optional, depending on applications.

An example of form without the "Pillar" level.

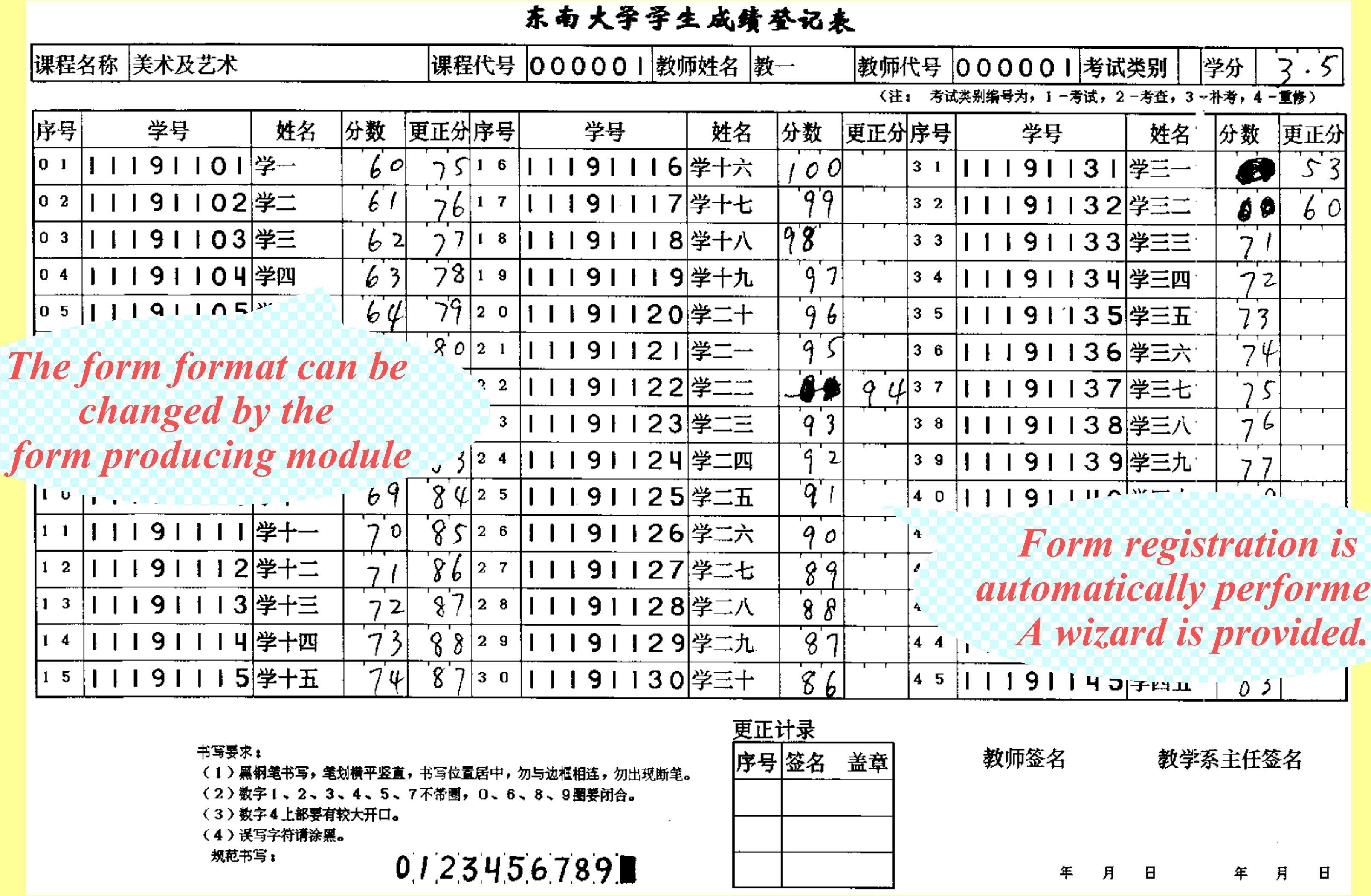




An example of form without the "Pillar" level.

2000-Jan-10

Sample Form



东南大学学生成绩登记表

课程名称	美术及艺术	课程代号	000001	教师姓名	教一	教师代号	000001	考试类别		学分	3.5

（注： 考试类别编号为，1－考试，2－考查，3－补考，4－重修）

序号	学号	姓名	分数	更正分	序号	学号	姓名	分数	更正分	序号	学号	姓名	分数	更正分
01	11191101	学一	60	75	16	11191116	学十六	100		31	11191131	学三一	[redacted]	53
02	11191102	学二	61	76	17	11191117	学十七	99		32	11191132	学三二	[redacted]	60
03	11191103	学三	62	77	18	11191118	学十八	98		33	11191133	学三三	71	
04	11191104	学四	63	78	19	11191119	学十九	97		34	11191134	学三四	72	
05	11191105	[illegible]	64	79	20	11191120	学二十	96		35	11191135	学三五	73	
[illegible]	[illegible]	[illegible]	[illegible]	80	21	11191121	学二一	95		36	11191136	学三六	74	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	22	11191122	学二二	[redacted]	94	37	11191137	学三七	75	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23	11191123	学二三	93		38	11191138	学三八	76	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	24	11191124	学二四	92		39	11191139	学三九	77	
[illegible]	[illegible]	[illegible]	69	84	25	11191125	学二五	91		40	[illegible]	[illegible]	[illegible]	
11	11191111	学十一	70	85	26	11191126	学二六	90		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	11191112	学十二	71	86	27	11191127	学二七	89		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	11191113	学十三	72	87	28	11191128	学二八	88		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	11191114	学十四	73	88	29	11191129	学二九	87		44	[illegible]	[illegible]	[illegible]	[illegible]
15	11191115	学十五	74	87	30	11191130	学三十	86		45	11191145	[illegible]	[illegible]	

书写要求：
（1）黑钢笔书写，笔划横平竖直，书写位置居中，勿与边框相连，勿出现断笔。
（2）数字1、2、3、4、5、7不带圈，0、6、8、9圈要闭合。
（3）数字4上部要有较大开口。
（4）误写字符请涂黑。
规范书写： 0123456789█

更正计录

序号	签名	盖章

教师签名　　年　月　日

教学系主任签名　　年　月　日

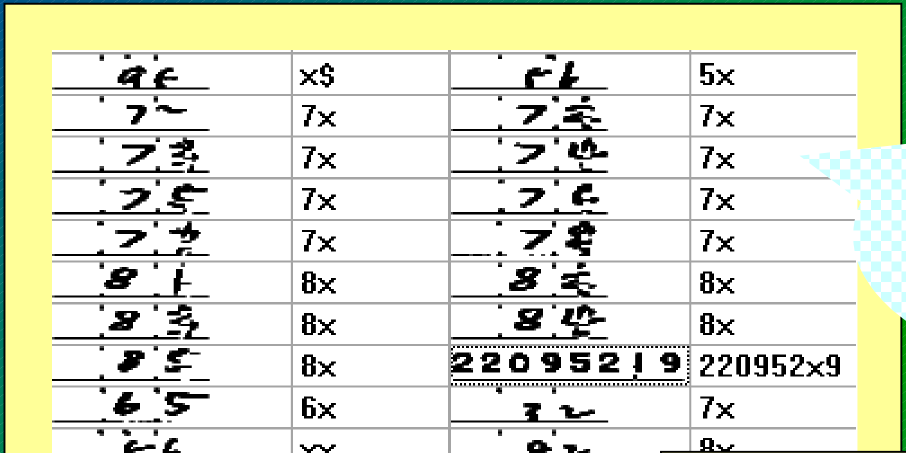
The form format can be changed by the form producing module

Form registration is automatically performed. A wizard is provided.

Handwritten Character Recognition

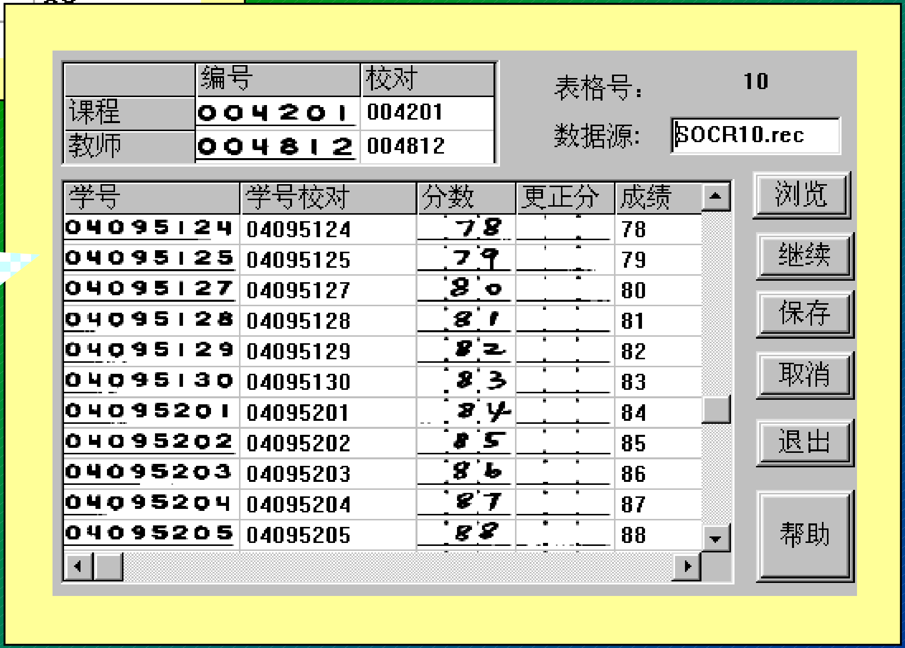
- The OCR/OMR module of SOCR can attain about **99.5%** recognition accuracy for loosely constrained (handprinted) handwritten numerals and symbols.
- "Loosely constrained" means that all the characters should not be indeterminate for classification. The constraints imposed on handwritten numerals include:
  - '0', '6', '8', '9' have closed circles,
  - '1', '2', '3', '4', '5', '7' do not have circles,
  - '4' does not close the upper part.
- The recognition engine of SOCR is based on a hybrid model, which combines both traditional methods and neural recognizers. Doubtful characters are sent to the rejection processing module.

Rejection Processing and Data Verification



	x$		5x
	7x		7x
	7x		7x
	7x		7x
	7x		7x
	8x		8x
	8x		8x
	8x	220952!9	220952x9
	6x		7x
	xx		8x
	x4		

Manual tool is provided for rejection correction. All the corrected characters can be distributed back to their original forms automatically.



	编号	校对
课程	004201	004201
教师	004812	004812

表格号： 10

数据源： SOCR10.rec

学号	学号校对	分数	更正分	成绩
04095124	04095124	78		78
04095125	04095125	79		79
04095127	04095127	80		80
04095128	04095128	81		81
04095129	04095129	82		82
04095130	04095130	83		83
04095201	04095201	84		84
04095202	04095202	85		85
04095203	04095203	86		86
04095204	04095204	87		87
04095205	04095205	88		88

Different data verification strategies are used to examine the (logical) correctness of recognized results. A pair examination wizard is provided.